CLAIMS

What is claimed is:

1. A system that facilitates spell checking, comprising:

a component that receives input data containing text; and

a spell checking component that identifies a set of potentially misspelled substrings in the text and proposes at least one alternative spelling for the substring set based on at least one query log; the query log comprising data utilized by users to query a data collection over a time frame.

2. The system of claim 1, the spell checking component further utilizes user-dependent information in proposing at least one alternative spelling.

3. The system of claim 1, the alternative spelling for the substring set is further based on at least one trusted lexicon; the trusted lexicon comprising at least one selected from the group consisting of a trusted lexicon with content and a trusted lexicon without content.

4. The system of claim 3, the spell checking component further employs a list of stop words; the list of stop words comprising at least one selected from the group consisting of a list of stop words with content and a list of stop words without content.

5. The system of claim 4, the list of stop words with content comprising a stop word list containing high frequency words and function words and their frequent misspellings.

6. The system of claim 4, the spell checking component employs an iterative process to search a space of alternative spellings.

7. The system of claim 6, the spell checking component employs, at least in part, heuristics to impose restrictions on a search space utilized to determine a proposed alternative spelling.

8. The system of claim 7, the heuristics utilize, at least in part, at least one fringe to limit the search space.

9. The system of claim 4, the query log comprising a histogram of queries asked over a time frame.

10. The system of claim 9, the histogram of asked queries relates to a subset of the users; the subset comprising at least one user.

11. The system of claim 9, the query log resides on a server computer.

12. The system of claim 9, the query log resides on a client computer.

13. The system of claim 9, the spell checking component utilizes substring occurrence and co-occurrence statistics from at least one query log.

14. The system of claim 13, a substring comprising at least one selected from the group consisting of an entry in at least one trusted lexicon, an entry in a stop word list, and a sequence of characters without a pre-defined set of delimiter characters.

15. The system of claim 13, the substring co-occurrence statistics comprising substring bigram counts; a substring bigram comprising a pair of substrings in a text.

16. The system of claim 15, the substring bigram comprising a pair of adjacent substrings in a text.

17. The system of claim 16, the substring co-occurrence statistics for the list of stop words with content further comprising a substring bigram with stop-word-sequence-skipping counts.

18. The system of claim 13, the substring occurrence and co-occurrence statistics from the query log are stored in a same searchable data structure.

19. The system of claim 18, the data structure comprising a trie.

20. The system of claim 18 handles concatenated *and/or* split substrings in a same manner as it handles individual substrings.

21. The system of claim 20, the spell checking component generates a set of alternative spellings that are substrings in at least one selected from the group consisting of at least one query log and at least one lexicon.

22. The system of claim 21, the set of alternative spellings comprising a set of alternative spellings determined *via* an iterative correction process.

23. The system of claim 22, the iterative correction process comprising a plurality of iterations that change at least one substring to another substring as an alternative spelling; the iterative correction process halts when all possible alternative spellings are less appropriate than a current set of alternative spellings.

24. The system of claim 23, the alternative spellings and their appropriateness are computed based on a probabilistic string distance and a statistical context model.

25. The system of claim 24, the probabilistic string distance comprising a modified context-dependent weighted Damerau-Levenshtein edit function that allows insertion, deletion, substitution, transposition, and long-distance movement of characters as point changes.

26. The system of claim 24, in each iteration, the set of alternative spellings for a substring is generated utilizing a searchable substring data structure extracted from at least one query log and at least one trusted lexicon.

27. The system of claim 26, in each iteration, the set of alternative spellings for each substring is restricted to within a probabilistic distance δ from an input substring; the restriction is imposed within each iteration without limiting the iterative correction process as a whole.

28. The system of claim 27, in each iteration, the iterative correction process searches for an optimum set of alternative spellings *via* utilization of a statistical context model.

29. The system of claim 28, the statistical context model comprising substring occurrence and co-occurrence statistics extracted from at least one query log.

30. The system of claim 29, a Viterbi search is employed to facilitate in determining the optimum set of alternative spellings according to the context model in each iteration.

31. The system of claim 30, the Viterbi search can employ fringes to restrict a search for alternative spellings in an iteration such that for every pair of adjacent substrings, if any of the substrings is in at least one trusted lexicon, then only one of the substrings is allowed to change in that iteration.

32. A method of facilitating spell checking, comprising:

receiving input data containing text;

identifying a set of potentially misspelled substrings in the text; and

proposing at least one alternative spelling for the substring set based on at least one query log; the query log comprising data utilized by users to query a data collection over a time frame.

33. The method of claim 32, the alternative spelling for the substring set is further based on at least one trusted lexicon; the trusted lexicon comprising at least one selected from the group consisting of a trusted lexicon with content and a trusted lexicon without content.

34. The method of claim 33, further comprising:

employing, at least in part, a list of stop words to facilitate in determining at least one alternative spelling; the list of stop words comprising at least one selected from the group consisting of a list of stop words with content and a list of stop words without content;

utilizing substring occurrence and co-occurrence statistics from at least one query log; the query log comprising a histogram of queries asked over a time frame and the substring occurrence and co-occurrence statistics from the query log are stored in a same searchable data structure;

handling concatenated *and/or* split substrings in a same manner as handling individual substrings; and

generating a set of alternative spellings that are substrings in at least one selected from the group consisting of at least one query log and at least one lexicon, the set of alternative spellings comprising a set of alternative spellings determined *via* an iterative correction process.

35. The method of claim 34, the iterative correction process comprising:

changing at least one substring to another substring as an alternative spelling; and

halting the iterative correction process when all possible alternative spellings are less appropriate than a current set of alternative spellings; the alternative spellings and their appropriateness are computed based on a probabilistic string distance and a statistical context model.

36. The method of claim 35, further comprising in each iteration of the iterative correction process:

utilizing a searchable substring data structure extracted from at least one query log and at least one trusted lexicon to generate the set of alternative spellings for a substring;

restricting the set of alternative spellings for each substring to within a probabilistic distance δ from an input substring; the restriction being imposed within each iteration without limiting the iterative correction process as a whole; and

searching for an optimum set of alternative spellings *via* utilization of a statistical context model; the statistical context model comprising substring occurrence and co-occurrence statistics extracted from at least one query log.

37. The method of claim 36, further comprising:

employing a Viterbi search to facilitate in determining the optimum set of alternative spellings according to the context model in each iteration; the Viterbi search can employ fringes to restrict a search for alternative spellings in an iteration such that for every pair of adjacent substrings, if any of the substrings is in at least one trusted lexicon, then only one of the substrings is allowed to change in that iteration.

38. A system that facilitates spell checking queries to a search engine, comprising:

means for receiving input data containing text; and

means for identifying a set of potentially misspelled substrings in the text and proposing at least one alternative spelling for the substring set based on at least one query log; the query log comprising data utilized by users to query a data collection over a time frame.

39. A data packet, transmitted between two or more computer components, that facilitates spell checking, the data packet comprising, at least in part, information relating to a spell checking system that provides, at least in part, at least one alternative spelling for a string set based, at least in part, on at least one query log.

40. A computer readable medium having stored thereon computer executable components of the system of claim 1.

41. A device employing the method of claim 32 comprising at least one selected from the group consisting of a computer, a server, and a handheld electronic device.

42. A device employing the system of claim 1 comprising at least one selected from the group consisting of a computer, a server, and a handheld electronic device.